QUANTARIUM Data & Analytics

Data Delivery Integration Document for Open Lien Data Set

Version 1.1e
Updated December, 2022

Document Control

Change Record

Date	*Author*	*Version*	*Change Description*
February, 2019	Ashish Bathini	1.0	New
January, 2020	Ashish Bathini	1.1a	Various Updates
April, 2020	Margaret Li	1.1b	Minor Update to Section 2.4.3
August, 2020	Margaret Li	1.1c	Added Section on Ingestion of Text Fiels
June, 2021	Margaret Li	1.1d	Added Guidance on FTP Transfer in Section 2.7
April 4, 2022	Margaret Li	1.1d	More guidance on FTP Transfer in Section 2.7
December 19, 2022	Margaret Li	1.1e	More details for Section 2.6 and Section 2.7

Contents

1 INTRODUCTION

1.1 Purpose

The purpose of this document is to describe the Quantarium Open Lien data asset files and the best practices for the data load recommended by the Quantarium Data Delivery Team.

Open Lien – a Quantarium managed data set

This is a managed data set which the Quantarium team has derived from its Public Records data assets. By doing deep, detailed cross reference, data cleanup and analysis, this data set contains property records data (including residential, commercial, and vacant land) and active mortgage (Open / Voluntary Lien) information associated with those properties.

1.2 Quantarium Contacts for Data Questions

- **Support Email:** Support@Quantarium.com
- **Data File Licensing Account Manager:** Darlene Davis
  - **Email:** Darlene.Davis@quantarium.com
  - **Direct Phone:** 760-587-1216

2 Fulfillment Delivery Requirements

2.1 Geographic Coverage

The "Quantarium_Data_Asset_Coverage_Report_Public - ***Date***.xlsx" file (available on request to the Data File Licensing Account Manager) contains coverage information for the underlying Public Records and the jurisdictions covered by Quantarium. The ***Date*** in the filename indicates the latest year/month/day that the coverage information is available.

2.2 Data File Format and Naming

2.2.1 File Name and Format Details

1. Bulk Files are tab delimited, with field headers. The layout of this data is described in an Excel file named **Quantarium_Open_Lien_Data_File_License_RLO_YYYYMMDD_Public.xlsx** where **YYYYMMDD** indicates the date of publication of that Record Layout.

2. Inside the Record layout Excel document, there are various tabs. The record layout guide resides in the tab labeled "OpenLien".

   The rows in the OpenLien tab are organized by: Field #, the Data Category of that field (e.g. Property ID, Ownership), the (suggested) Field Display Name if it were to be displayed in an User Interface, the associated Header Name[1] in the actual bulk file, the maximum length of that field, the (suggested) data type in the database, the Data Format example, the look up tab and location if the data field returns codes, and the associated description for the field.

   In addition, the Record Layout file has the following tabs:

Record Layout Tab	Information
FIPS_list	Maps FIPS Code to Postal State and County Names
CodeXlate	Maps Codes in actual data to their corresponding human readable strings
GeoMatch Code	Mappings for Property Address Match_Code fields
GeoLocationCode	Mappings for Property Address Location_Code fields

3. Bulk File Naming Convention:
   Quantarium_OpenLien_YYYYMMDD_SeqNo.zip

   An example of this naming convention would be: Quantarium_OpenLien_20191012_0001.zip

   Note that while the underlying data files are in Tab delimited (TSV) format, they are compressed and delivered as ZIP files. Each TSV file is split up after reaching 5 million records. In the above example, the bulk file named Quantarium_OpenLien_20191012_0001.TSV is therefore located in the zipped file named Quantarium_OpenLien_20191012_0001.zip. The **YYYYMMDD** refers to the date when the data in the bulk file was delivered.

[1] From time to time, Quantarium may need to change the header names for consistency and alignment. We will provide notification in advance.

2.3 Description of Folders and Files

The full data set resides in an FTP folder of that name, appended with the version (e.g. V2.6). There is also a folder with the weekly deltas or updates for that version. For example:

2.3.1 Full Set

For the Open Lien full data set, this folder ("Open Lien Full Set V2.6) contains the full nationwide data set. There are 3 different types of files in this folder:

a. **Data files:** these contain the records for the Open Lien Data Set. The naming convention for these files is **Quantarium_OpenLien_YYYYMMDD_SeqNo.Zip** as described above.

b. **Control file:** this contains summary statistics (see more in Section **2.5** Metadata Files Description) and, more importantly, its presence is an indicator that the data set is completely delivered and is ready for downloading by our customers. The naming convention of this file is **Quantarium_OpenLien_ControlFile_YYYMMDD.TSV**

c. **Full Counts file:** this contains coverage statistics and fill counts for all the fields in the Data Set. The naming convention of this file is **Quantarium_Open_Lien_Full_Counts_YYYMMDD_00001.TSV**

In summary, below are the file name templates for Open Lien Data Sets:

- Quantarium_OpenLien_YYYYMMDD_SeqNo.zip
- Quantarium_ OpenLien _ControlFile_YYYYMMDD.TSV
- Quantarium_ Open_Lien _Full_Counts_YYYYMMDD_00001.CSV

NOTES:
The Record Layout document also resides in the Full Set folder, for reference.
Delete files are not applicable and thus not delivered for the Full set.
Full sets are *usually* generated for each Quarter – the actual delivery date in the quarter will be announced to customers who are looking for a full refresh.
The Full Counts file is available for the Full Set only and not for Delta files.

2.3.2 Weekly Deltas

For the Open Lien Data Set, the Deltas folder (Weekly Deltas V2.6) contains the nationwide weekly changes to the data set. Changes comprise of deleted records and new and/or changed records. These are available each Friday prior to or by 9AM Eastern Time.

Delete files identify those records that need to be removed. The naming convention for these files is: **Quantarium_OpenLien_Delta_Delete_YYYMMDD_seqNo.TSV** (which is packaged inside a zipped file).

Update files containing new and/or changed records. The naming convention for these files is: **Quantarium_OpenLien_Update_YYYMMDD_seqNo.TSV** (which is packaged inside a zipped file).

While unlikely, it is possible for a Delete file or an Update file to have zero records (but the header row will still be present). This means no updates or no deletes were received that Week.

For Open Lien weekly delivery, we deliver following delta files:

- **Quantarium_OpenLien_Update_YYYYMMDD_SeqNo.zip**
- **Quantarium_OpenLien_Delta_Delete_YYYYMMDD_SeqNo.zip**

In addition, we deliver a Control File for that week (more on this in section **2.4.2** Weekly Update Processing):

Quantarium_OpenLien_ControlFile_YYYYMMDD.TSV

All the delta files for a given week would have the same (date) value for **YYYYMMDD**.

2.4 Process Description

2.4.1 Full Set Processing

We are providing sample SQL scripts (see OpenLien.sql file) to help with the Extraction, Transformation and Loading (ETL) process. They include creating the prestaging "load" table, "permanent" table as well as the stored procedure for converting/loading data from "load" to "permanent" table.

The "load" table is a temporary (staging) storage where all fields are defined as **[VARCHAR](500)** type and loaded directly from the tab-delimited text/ASCII files; the "permanent" table has each field matching the actual type from the Data Dictionary. We recommend the best practice steps below for "Full Set" loading.

Note that these SQL scripts are used with a Microsoft SQL Server database environment so it's possible these need adjustments according to the customer's specific SQL database server environment.

For Full-Set processing it is critical to *first* delete ALL existing data in the database tables and re-load these full set TSV files encapsulated in these zipped files below:

Quantarium_Openlien_YYYYMMDD_SeqNo.zip

into the corresponding Open Lien table, by taking the following steps in the ETL process:

1. Truncate Open Lien "load" table (the "load" table is referenced as "`Quantarium_Staging_OpenLienData`" in the provided SQL script).
2. Loop through each of the zipped files, unzip and load data into the "load" table.
3. Truncate Open Lien "permanent" table (the "permanent" table is referenced as "`Quantarium_OpenLienData`" in the provided SQL script).
4. Run the SQL script (stored procedure) to load data from "load" to "permanent" table (this is the provided SQL INSERT script).

2.4.2 Weekly Update Processing

This section is applicable for those who elect to keep the data set up to date on a week-by-week basis.

Once the Full Set of data files is loaded, the data sets can be kept up to date each week using incremental delete/update weekly files from the Weekly Deltas folder. It's important that they are loaded in order by date (YYYYMMDD), and the sequence mentioned below for each date.

For each weekly data set, always ensure that the associated Control File is present prior to downloading as its presence indicates the delivery is complete for that week's data set.

The files referenced below are the TSV files which reside in the zipped files and need to be processed in this sequence:

1. Use the **Quantarium_OpenLien_Delta_Delete_YYYYMMDD_SeqNo.tsv** to delete all applicable records from Open Lien table with the Quantarium_Internal_PID values in this file.

2. Load all records from **Quantarium_OpenLien_Update_YYYYMMDD_SeqNo.tsv** file into the Open Lien table.

2.4.3 Refresh Alternatives and Considerations

As already mentioned, a "Full Set" is refreshed every quarter, while the weekly updates are delivered once a week on or prior to Fridays 9AM Eastern Time.

A new customer must first load the full Open Lien data set. After that, if a customer has contracted to obtain weekly updates, then all the available deltas can be ingested in a manner described in the prior section.

When the next quarterly full refresh becomes available, an existing customer can:

- Reload (again) the "Full Set" and thus do a clean fresh start; *OR*
- Continue to load the weekly delta, without reloading the Full Set.

The advantage of the Reload is that all the managed fields in the data set (for example, the Financing and Valuation fields whose values are provided based on Quantarium's value-added analytics) will be refreshed even if the public records data of the properties did not change because Quantarium is using the most recent interest rate and environmental data to recalculate these fields.

On the other hand, customers can choose to not reload and just continue to load the weekly data. This has the advantage that the incremental load is faster since the weekly delta is significantly smaller than the "Full Set", but the managed fields in the data set will not get refreshed. Many counties (but not all) do ultimately fully refresh their public records data; as these refreshes occur, these managed field values will get recalculated, triggered by the refresh. It is important to note some counties may or may not refresh in any given calendar year. As a result, the managed field for static properties may not get updated. During relatively stable housing markets, these managed field values would likely not diverge much; to address any such concern, the full refresh can be leveraged.

Please keep in mind of **this scenario**:

- It is possible for a weekly update to contain a subset of records already delivered. This situation occurs usually when the full (nationwide) data set has just become available and then the subsequent weekly refresh would contain a small portion of records that overlaps the full data set. This occurs to ensure customers who do not wish to consume the full data set will still get the full week's update since the prior week's delivery, because the full refresh usually occurs in between two weekly deliveries.
- Very occasionally, there can be an overlap of a day's processing for the weekly updates due to rare and unique timing of the file creation in the system.

To ensure everything goes smoothly, we suggest the update file processing should encompass (upon encountering an existing record) deleting the existing record and replacing it with the new record (for the same PropertyID identifier) regardless of when it was received. By doing this, this process would delete the old record and add the new record, avoiding conflict.

For a customer ingestion situation where all the updates are stored in one table, and the combination of PropertyID and File_Creation_Date is used as a unique key, then first deleting an old record prior to inserting the new record with that exact PropertyID will guarantee against conflicts. If there is a need to be truly unique in this example, then it is best to add a "file received date" to the unique key in order to isolate the individual records.

2.5 Metadata Files Description

For data reconciling purposes, there are metadata files (Control Files) provided. The control file for the Open Lien data set has a record count for each FIPS codes where data is provided, and the total number of all the provided records should match the total record count in the Control Files:

Full Set V2.6 Quantarium_OpenLien_ControlFile_YYYYMMDD.TSV

Weekly Deltas V2.6\Quantarium_OpenLien_ControlFile_Update_YYYYMMDD.TSV

Data elements in control files (tab delimited format with field headers) are::

" RecId "
" FIPS_Code "
" State_Code "
" County_Name "
" Assessment_RecordCount"
" Mortgage_MinDate"
" Mortgage_MaxDate "

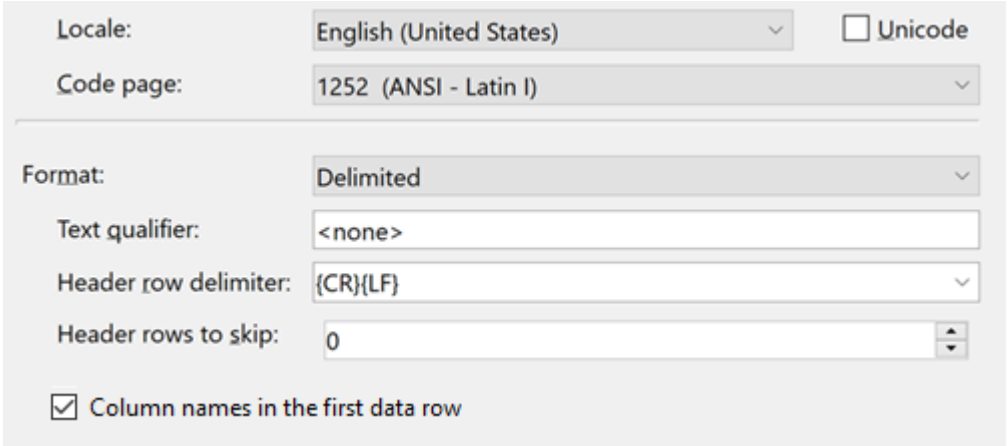
The Assessment_RecordCount field provide the record count for each FIPS geography.

2.6 Ingestion of Text Fields

Sometimes the ingestion program's tab-separated values parser fails over a row and this results in what *appears* to be missing tab(s) in the data, thus causing one or more records to not process successfully. This issue is usually triggered by a combination of:

- Presence of double-quote characters in character fields (we see this most often with the Legal_Brief_Description fields in the Assessment data set) **and**,
- The default setting in (quite a few) data loading programs (Excel included) where the text qualifier is set to double-quotes. The double-quotes in the data fields then could unintentionally infer additional fields, conflicting with the Tab Separated Values (TSV) definition for those records.

To boost the resilience of data loading program, please be sure to set the Locale to US English and Code Page to ANSI and ensure the Text qualifier property is not set to any specific character. For example, for Microsoft SQL Server, use the settings shown below:



2.7 Delivery Method

All files will be delivered via a secure FTP server: quantariumftp.hostedftp.com. An account will be provided to you by the Quantarium Account Manager.

After every Full Quarterly Refresh for the active version, the previous Full Set will be moved to a folder named "Full Set Vn.m - <*Previous Quarter+Year*>", where n.m indicates the active version number Only the previous quarter's full refresh will be available. For example, if the currently active Open Lien version is V2.6, and if the latest full refresh was generated in Q1 2021, and the prior full refresh was generated in Q4 2020, then existing customers who wish access to both will see "Full Set V2.6 – Q42020" and "Full Set V2.6" folders. Customers on-boarded in the current quarter (e.g. Q1 2020) will only see the (latest) "Full Set V2.6" folder.

The weekly deltas for version V2.6 can always be found in the "Weekly Deltas V2.6" folder. This folder contains weekly updates that may go back farther than the date for the most recent Full Set, so the starting point of the weekly deltas is determined by the date associated with the latest ingested Full Set. For any given Full Set, the first applicable weekly update is the one with the closest YYYYMMDD value that is newer than the YYYYMMDD date of the ingested Full Set.

PLEASE NOTE:
We STRONGLY ADVISE that for the Open Lien Full Set, you first download ***all*** the Full Set files from the Quantarium FTP to your local file store before you begin processing the individual files. Multiple / repeated downloads of the same files (each of which can be quite large) can exhaust the Quantarium FTP download bandwidth and thus impact other Quantarium customers.

We also advise that even for the smaller daily update files, please do not repeatedly download them as this action can accumulate to cause bandwidth impact. It is best to download files to a local staging file share for downstream processing. **Note also** that auto-resuming on a failed FTP connection can result in full transfers, not just requests; and this translates to Quantarium' FTP seeing multiple back-to-back full transfers of the same file that take up bandwidth. Please ensure the download error handling accounts for this scenario appropriately.\

It is important to just download the newest update files and not download all the files in the **Weekly Deltas V2.6** folder from prior weeks' deliveries which have already been downloaded, to achieve optimal transfer bandwidth and data storage. One way to do this is to maintain a text file (let's call this DownloadedFiles) where the names of previously downloaded filenames are kept. Then the logic to download new files in a Quantarium FTP folder is:

1. Instantiate an FTP client:
   ```
   SftpClient sftpClient = new SftpClient(_ftpHost, _ftpUser, _ftpPdw)
   ```
2. Fetch the name of all files in the Quantarium FTP folder (let's call this list ftpFiles):
   ```
   ftpFiles = (List<SftpFile>)sftpClient.ListDirectory(_ftpFolder);
   ```
3. Iterate through ftpFiles. If a filename is not in DownloadedFiles list, then initiate FTP download for that file.
   ```
   using (FileStream fs = new FileStream(destinationFile, FileMode.Create))
   {
           sftpClient.DownloadFile(ftpfiles.FullName, fs);
   }
   ```
4. After each file is successfully downloaded, append the newly downloaded filename to DownloadedFiles.

There are other ways to do this of course, such as remember the date/time when the prior download took place and download all files written since that time.